FIFE user activity mitigation policy, March 2020

Fermilab provides shared computing resources spanning multiple categories of computing, networking, storage, and monitoring. These complex and distributed services can be affected adversely by unusual or unexpected activities from users of the systems. This policy describes the mitigation strategy for these events and the path of escalation and remediation that will be taken to restore the computing services to normal operations.

Sometimes users will submit jobs that are wasteful (request much larger resources than needed), highly inefficient (low efficiency thresholds are defined elsewhere), or cause a degradation or outage to a shared resource (swamping worker node I/O, causing a large transfer queue on dCache pools, or overloading a database are some examples). *Note these measures should not be implemented in cases where there is a known service outage or degradation not caused by the user that could be impacting their workflow(s).*

Definitions

“Experiment support” means the individual(s) that an experiment designates to work with users and service providers to diagnose issues with jobs. That may include the CS liaison in some cases but it may be others as well. It is the experiment’s responsibility to ensure that it is populated with the right people to troubleshoot and debug jobs. There should be a single point of interface (SNOW) between the service providers and this support team. The experiment is free to use any means it wishes to debug, solve problems, and document work (SNOW, Slack, internal mailing lists, etc.) internally. Communication with the service providers, however, should be via the specified interface so that both the experiment and service providers have access to the record.

“Contacting the user” means using the email sent to the user’s official address in the FNAL telephone book (https://www-tele.fnal.gov) on SNOW tickets. That is not required to be a Fermilab (i.e. username@fnal.gov) email. However, users are responsible for ensuring that the address listed there is one that they check frequently.

“DCS Primary” is the weekly on-call person who initially deals with SNOW tickets assigned to Distributed Computing Support.

Situations covered by the policy

Situations that will elicit a response from SCD according to the policy include, but are not limited to:

1. Workflows that cause a degradation, outage, or denial of service to a shared resource. These include overloading a database, causing a large transfer queue on one or more dCache pools, excessive disk IO that impacts the entire machine, excessive memory usage beyond the job's request that results in entire glideins being killed.
2. A failure rate above 50 percent on jobs that result in a waste of over 1,000 slot-weighted hours of wall time in a 24-hour period. Jobs finishing with a non-zero exit code are considered failed jobs, even if they are successful from the experiment's point of view.
3. Analysis jobs that consume more than 1,000 slot-weighted hours of wall time and are below the FIFE efficiency policy thresholds (defined elsewhere) in CPU, disk, or memory efficiency (whether or not the jobs are successful) for 3 or more days in a 7-day period (including weekends and/or laboratory holidays).
4. Overall production output that is in violation of the efficiency policy for more than 24 hours and consumes more than 1,000 slot-weighted hours.
5. Any job that violates the Fermilab security policy in any way is subject to immediate removal, and the user who submitted the job is subject to the consequences outlined in the policy.

For the purposes of this policy, the originator of the activities is classified into two general classes - "Official Experiment Production Activities" and "General Experiment Activity" and treated separately due to the expertise and communication channels that are available in each case. The fundamental distinction is that production activities will be ramped down to a sustainable level, while individual users' jobs will be held as needed.

Levels of response

If there is an issue that is not causing a significant impact to shared resources, notify user(s) and/or production teams and experiment support of the problem and explain that they need to take action to mitigate it. The notification will be in the form of a SNOW ticket assigned to the experiment support group with the user(s) on the watch list. **If the impact is significant and cannot wait for a reply, proceed to the next steps without waiting for a response.** The DCS primary should determine whether the impact is significant in consultation with experts from the affected service(s). If there is not a significant ongoing impact **but there is no response from the user(s) after one business day, proceed to the steps below.** There are different actions for production and user analysis jobs.

Official Experiment Production Activities:

1. A Service ticket reporting the incident will be opened through the FNAL Service portal (https://fermi.service-now.com/) using a custom form assigning to FIFE support.
    a. If details regarding the degradation (source, impacted systems, etc…) are available, they will be included in the initial ticket.

    b. The ticket will be handled by the FIFE Support on-call mechanism and in accordance with the service level agreement for the affected services

2. The ticket will include the following workflow steps
    a. Create an approval task for SCD management so that the following actions can proceed.
    b. Set the quota for the corresponding accounting group to affect a ramp down in the number of concurrent applications being run by the experiment.This may be an iterative process to determine an adequate value to restore operations.
    c. Restrict the accounting group in the batch systems [through setting of the GROUP_ACCEPT_SURPLUS flag to FALSE] to prevent an inadvertent reoccurrence of the situation.[1]
    d. Create a ticket in the Service Now (SNOW) system corresponding to investigation and resolution of the problem, assigning to the experiment-support team.
    e. Notify experiment spokespeople or otherwise designated points of contact about the problem and the remediation steps.

General Experiment Activity (User Analysis Jobs)

1. Hold the user's jobs in the accounting group in question also notifying experiment support in the process. The DCS Primary can use the FERRY API to do this; it will set the is_banned flag for the accounting group in question. This will also prevent additional jobs from starting. This hold would remain in place until the user satisfies the reinstatement policy. The primary should note this action in the ticket if one already exists for this situation. If there is no ticket, the DCS Primary can make a new one and assign it to experiment support. Most situations will not require any additional escalation.
2. Repeat violators and/or unresponsive users are subject to extended bans across all accounting groups of which the user is a member, and may involve senior management (both from SCD and the experiment) as needed to be sure that the behavior is corrected. A record would be made in this case (a SNOW ticket to experiment support) with all relevant parties from both SCD and the experiment on the watch list. The DCS primary should also send an initial email directly to the relevant senior management individuals so that they know to follow this particular ticket. The ban would remain in place until the user satisfies the reinstatement policy.

[1] Condor also has "concurrency limits". If there are problems with the "non-surplus" quota plan we could investigate setting up concurrency limits for each experiment. We actually have some in the system now and have used them in the past but it seems no one has used them for some time.

Reinstatement policy

The reinstatement process begins with users replying to the notification they received during the previous steps. The experiment support group should be included on all traffic to develop a remediation:

1. Following best practices, the experiment support teams should work with the user and/or production team to correct the part(s) of the workflow causing the problem (the remediation).
2. After creating the remediation the user should open a ticket asking for reinstatement for testing purposes. The user can then send a small-scale test of the corrected workflow (where small scale is defined as the smallest number of jobs necessary to verify good operation, usually 1-100 jobs). If the user immediately sends the full workflow without a test, they will be banned again and the process will restart.
3. After the remediation is successfully tested at a small scale, the user can resume larger submissions. However, SCD personnel may require additional scale testing before full reinstatement, as some problems only appear in high-concurrency situations.

Follow Up, Documentation, Debriefing

For emergent situations: after the situation is resolved and services have been restored to normal operating conditions, computing sector personnel and experiment representatives will work to both document and disseminate information relating to the incident. This work will include:

- Resolution and closing of all SNOW tickets relating to the incident
- Documentation of the details of the cause of the degradation and of the steps taken to solve the incident as well as any changes to experiment workflows or applications that were made to prevent future incidents
  - This documentation should be recorded in both the CS documentation systems (by CS personnel), and in experiment-specific documentation systems (i.e. experiment specific computing howto's, wiki instructions, officially maintained code bases) by experiment representatives.
  - A report on the incident will be prepared and presented at the next FIFE meeting so that other experiments can be alerted to the incident and how to prevent similar incidents.
- The incident will be reported to SCD management through normal reporting channels.

For non-emergent situations (those not causing a service degradation), work includes closing the tickets and transferring any lessons learned to the standard best practices documentation, along with any other communication with experiments deemed necessary to disseminate the information. This step would typically be an email to the cs-liaison mailing list.